The Secret of Overlapping Tokens in Question Generation

Anonymous ACL submission

Abstract

This document is a supplement to the general instructions for *ACL authors. It contains instructions for using the LaTeX style files for ACL conferences. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used both for papers submitted for review and for final versions of accepted papers.

1 Introduction

Question generation (QG) in reading comprehension (RC) topic is the task of generating natural question given a pair of passage and answer. The answer is a span of passage that aim to be answered by the generated question. The advances of producing synthetical questions shall aid the data collection process for low resources languages as many of large-scale RC datasets (Rajpurkar et al., 2016 ; Joshi et al., 2017) requires crowdworkers to gather the list of questions that is costly to conduct. Other than that, it may also benefit the educational area to produce questions for RC materials (Heilman and Smith, 2010).

To put into another perspective, QG poses similar workflow with the summarization task which generally yield a single sentence by understanding and identifying salient information from multiple sentences. Therefore, we utilize a pre-trained summarization model, Pegasus (Zhang et al., 2020) to maximize the capability of capturing relevant information from the passage. However, there are three phenomena that differentiate QG with summarization that we explicitly highlight in this study. These phenomena were found in SQuAD dataset (Rajpurkar et al., 2016) and we further leverage those to design a more robust QG system.

First, the natural trait of asking question in RC is to sought out for answer that assumed to be located somewhere in the passage. As the answer itself unknown by the questioner, the likelihood for the entire answer tokens to be present in the target question is tiny. For example, one of the questions in SQuAD, "which university has its origins in a school dealing with medicine and surgery?" has an answer, "newcastle university". In this particular instance, it is unlikely to include both tokens "newcastle" and "university" to the question and remain coherent even with multiple way of rephrasing.

The previously proposed work (Du et al., 2017) utilize deep neural network (NQG) without any preventive approach towards this phenomenon. Thus the model tends to include the answer tokens to the output sequence. To mitigate this issue, (Kim et al., 2018 ; Yui et al., 2019) mask the answer with a special token (for e.g., <ANS>) and concatenate the answer to the end of the passage. Although this approach enable to reduce significant amount of output sequences that have answer tokens included, there is no guarantee for the model to not sample the answer tokens during decoding. To improve this approach, we proposed a decoding technique to control the probability of sampling answer tokens with adjustable temperature.

Second, we consider the role of interrogative words are integral for QG in RC because the incorrect usage could easily break the semantic structure of a question. For instance, replacing the token "who" with other interrogative tokens (e.g., what, when, which, or how) for the question, "who is the american singer that regarded as the king of pop in 1984?" will be semantically inappropriate. Aware by this, (Yui et al., 2019) extract the interrogative word directly from the target question as the upperbound as well as the surrounding token (e.g., “how many” or “in what year”) while (Kang et al., 2019) utilize transformer-based model (Vaswani et al., 2018) to classify the interrogative word. In this study, we follow (Kang et al., 2019) approach to produce the interrogative word as multi-task learning setting.

Finally, we called a token as alpha token, if it satisfied the following properties: 1) It is overlapped between passage and question. 2) It is non-stop words. We investigate this alpha token further in SQuAD dataset and found most of them appear in the same sentence with the answer tokens. Accordingly, we pull out this sentence and mask each token that satisfied alpha token properties. Furthermore, we explore BART (Lewis et al, 2020) that was pre-trained to reconstruct masked text and utilize this model to predict the masked alpha tokens. The predicted alpha tokens, then used as additional features along with the predicted interrogative word for the main QG model.

The approaches described above were tailored to deal with QG phenomena which successfully outperform the existing best models by large margin.

2 Engines

To produce a PDF file, pdfLATEX is strongly recommended (over original LATEX plus dvips+ps2pdf or dvipdf). XeLATEX also produces PDF files, and is especially suitable for text in non-Latin scripts.

3 Preamble

The first line of the file must be

```
\documentclass[11pt]{article}
```

To load the style file in the review version:

```
\usepackage[review]{acl}
```

For the final version, omit the `review` option:

```
\usepackage{acl}
```

To use Times Roman, put the following in the preamble:

```
\usepackage{times}
```

(Alternatives like txfonts or newtx are also acceptable.)

Please see the LATEX source of this document for comments on other packages that may be useful.

Set the title and author using `\title` and `\author`. Within the author list, format multiple authors using `\and` and `\And` and `\AND`; please see the LATEX source for examples.

By default, the box containing the title and author names is set to the minimum of 5 cm. If you need more space, include the following in the preamble:

```
\setlength\titlebox{<dim>}
```

where `<dim>` is replaced with a length. Do not set this length smaller than 5 cm.

Command	Output	Command	Output
`{\"a}`	ä	`{\c c}`	ç
`{\^e}`	ê	`{\u g}`	ğ
``{\`i}``	ì	`{\l}`	ł
`{\.I}`	İ	`{\~n}`	ñ
`{\o}`	ø	`{\H o}`	ő
`{\'u}`	ú	`{\v r}`	ř
`{\aa}`	å	`{\ss}`	ß

Table 1: Example commands for accented characters, to be used in, *e.g.*, BibTEX entries.

4 Document Body

4.1 Footnotes

Footnotes are inserted with the `\footnote` command.[1]

4.2 Tables and figures

See Table 1 for an example of a table and its caption. **Do not override the default caption sizes.**

4.3 Hyperlinks

Users of older versions of LATEX may encounter the following error during compilation:

```
\pdfendlink ended up in
different nesting level
than \pdfstartlink.
```

This happens when pdfLATEX is used and a citation splits across a page boundary. The best way to fix this is to upgrade LATEX to 2018-12-01 or later.

4.4 Citations

Table 2 shows the syntax supported by the style files. We encourage you to use the natbib styles. You can use the command `\citet` (cite in text) to get "author (year)" citations, like this citation to a paper by Gusfield (1997). You can use the command `\citep` (cite in parentheses) to get "(author, year)" citations (Gusfield, 1997). You can use the command `\citealp` (alternative cite without parentheses) to get "author, year" citations, which is useful for using citations within parentheses (e.g. Gusfield, 1997).

[1] This is a footnote.

Output	natbib command	Old ACL-style command
(Gusfield, 1997)	`\citep`	`\cite`
Gusfield, 1997	`\citealp`	no equivalent
Gusfield (1997)	`\citet`	`\newcite`
(1997)	`\citeyearpar`	`\shortcite`

Table 2: Citation commands supported by the style file. The style is based on the natbib package and supports all natbib citation commands. It also supports commands defined in previous ACL style files for compatibility.

4.5 References

The LaTeX and BibTeX style files provided roughly follow the American Psychological Association format. If your own bib file is named `custom.bib`, then placing the following before any appendices in your LaTeX file will generate the references section for you:

```
\bibliographystyle{acl_natbib}
\bibliography{custom}
```

You can obtain the complete ACL Anthology as a BibTeX file from `https://aclweb.org/anthology/anthology.bib.gz`. To include both the Anthology and your own .bib file, use the following instead of the above.

```
\bibliographystyle{acl_natbib}
\bibliography{anthology,custom}
```

Please see Section 5 for information on preparing BibTeX files.

4.6 Appendices

Use `\appendix` before any appendix section to switch the section numbering over to letters. See Appendix A for an example.

5 BibTeX Files

Unicode cannot be used in BibTeX entries, and some ways of typing special characters can disrupt BibTeX's alphabetization. The recommended way of typing special characters is shown in Table 1.

Please ensure that BibTeX records contain DOIs or URLs when possible, and for all the ACL materials that you reference. Use the `doi` field for DOIs and the `url` field for URLs. If a BibTeX entry has a URL or DOI field, the paper title in the references section will appear as a hyperlink to the paper, using the hyperref LaTeX package.

Acknowledgements

This document has been adapted by Steven Bethard, Ryan Cotterell and Rui Yan from the instructions for earlier ACL and NAACL proceedings, including those for ACL 2019 by Douwe Kiela and Ivan Vulić, NAACL 2019 by Stephanie Lukin and Alla Roskovskaya, ACL 2018 by Shay Cohen, Kevin Gimpel, and Wei Lu, NAACL 2018 by Margaret Mitchell and Stephanie Lukin, BibTeX suggestions for (NA)ACL 2017/2018 from Jason Eisner, ACL 2017 by Dan Gildea and Min-Yen Kan, NAACL 2017 by Margaret Mitchell, ACL 2012 by Maggie Li and Michael White, ACL 2010 by Jing-Shin Chang and Philipp Koehn, ACL 2008 by Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, ACL 2005 by Hwee Tou Ng and Kemal Oflazer, ACL 2002 by Eugene Charniak and Dekang Lin, and earlier ACL and EACL formats written by several people, including John Chen, Henry S. Thompson and Donald Walker. Additional elements were taken from the formatting instructions of the *International Joint Conference on Artificial Intelligence* and the *Conference on Computer Vision and Pattern Recognition*.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

A Example Appendix

This is an appendix.